5.2 A 16-Core RISC Microprocessor with Network Extensions

Vishnu Yalala, Derek Brasili, David Carlson, Adam Hughes, Anil Jain, Tim Kiszely, Kolar Kodandapani, Anand Varadharajan, Thucydides Xanthopoulos

Cavium Networks, Marlborough, MA

This 600MHz multi-core RISC processor is targeted for layer-4 through layer-7 network applications. It exhibits a high level of integration: 16 custom MIPS64 RISC cores, 1MB L2 cache, hardwired security engines, network function accelerators, and memory/network/bus controllers. The processor is designed for power efficiency, which is a key requirement for embedded applications [1]. Figure 5.2.1 shows the chip floorplan. Most of the silicon area is dedicated to the 16 RISC processors and the 1MB L2 cache. The remaining area is occupied by network coprocessors and physical interfaces. The chip interfaces consist of a 64b 133MHz PCI/PCIX, two 16Gb/s SPI4.2 or eight 1Gb/s RGMII, a 144b 800MHz DDR2, a 36b 600MHz low-latency DRAM interface in addition to miscellaneous and general-purpose I/Os. This 180M transistor 25W processor is fabricated in a 1.2V 0.13μm CMOS process with 9 layers of copper interconnect using FSG dielectric and C4 bumps. The performance is summarized in Fig. 5.2.2.

The chip contains dedicated network coprocessors for parsing, error checking, tagging, queuing, work scheduling, and managing the physical interfaces with support for IPv4 and IPv6 traffic up to 20Gb/s. For anti-virus, IDS and content processing applications, the chip contains dedicated deterministic finite automata (DFA) coprocessors to accelerate pattern and signature matching at up to 4Gb/s. The TCP acceleration coprocessor performs hardware-based packet synchronization, timer support, and buffer management to deliver 10Gb/s of full TCP termination. Finally, there is a 4Gb/s hardwired ZIP compression/ decompression accelerator. All these engines interface with 16 RISC cores and 1MB L2 cache through an I/O bridge.

Each RISC core can issue in-order two MIPS instructions per cycle from the 32kB 4-way set-associative virtual instruction cache. The execute unit consists of two pipelines: The first handles all instructions and the second only handles ALU/insert/extract/shift/move instructions. The memory section consists of an 8kB fully associative Dcache, a 2kB write buffer, and a 32-entry (64 page) unified translation look-aside buffers (TLBs). The multiplication/division unit in addition to supporting the standard MIPS instructions, implements a vectored IMUL instruction to accelerate modular exponentiation. This allows a 64b multiply and add to execute every cycle. Cryptographic operations are accelerated by dedicated units supporting different encryption methods: 3DES, AES, MD5, SHA1/256/512, and GF2. Finally, a dedicated unit accelerates cyclic redundancy checks and also implements counting leading zeros/ones and population-count instructions. Figure 5.2.3 shows a floorplan of each RISC core along with the pipeline flow diagram.

Power for an individual processor is 450mW @600MHz. This is achieved primarily through aggressive clock gating of all place-and-route and custom islands. Some blocks present natural exclusivity, which is exploited. A good example is the execution unit where for a given instruction only the ALU or the shifter need be enabled, not both. In other cases, the hardware enforces exclusivity to reduce peak power. For example, the vector IMUL instruction that consumes 80mW, will not execute when any of the crypto engines are in use. The issue- and memory-control sections, implemented in automated flow, have a substantial number of flops needed only for debug and privileged state. These flops are placed on a separate clock domain that is activated so rarely and can be ignored for power considerations. Taking into account the core, the L2 cache, and global clock generation and distribution, the power performance is approximately 2000 MIPS/W.

The 32kB Icache presented a number of opportunities for power reduction. The organization is four 8kB sets, with simultaneous tag and data lookup. Each set consists of 256 rows with 4-way column multiplexers to yield a 65b fetch result. Traditionally, the column multiplexers are interfaced to the bitlines prior to the sense amps. In this design, the sense amps are directly attached to the bitlines and the column multiplexers operate on the sense-amp outputs (Fig. 5.2.4). The stacked input devices improve the statistical sensing capability of the circuit for sub-100mV input differentials. Figure 5.2.5 compares the current sense amp with one single (non-stacked) input device of equivalent pulldown strength under V_t, W, and L variations. The proposed circuit requires 40mV less differential, which results in a substantial reduction (330ps) of the cache cycle time (120mV/ns bitline slew rate). The feedback keeper devices turn the sense amp into a static latching structure under changing data inputs. This is required because the Icache supports a fetch-under-fill, where the sense amps need to hold state while a write sequence is driving a full swing value on the bitlines. The power advantage of this approach is that a substantial portion of the Icache power is consumed driving the wordlines and bitlines. The incremental power of the additional sense amps is more than compensated by being able to fetch from the same 256b row merely by changing the column multiplexer selects. When executing sequential code, no full cache access is required until a 256b boundary is crossed. When crossing a 256b boundary, only one set is accessed, unless a 128B cache line is also crossed. For tight loops, set prediction is implemented allowing a single set access.

The 16 processors share a 1MB fully coherent L2 write-back cache. The interface to the cache consists of a 256b fill bus and a 128b store data bus. A separate command bus exists to send fill requests to the L2. The command and store buses are arbitrated by a single-cycle round-robin arbiter shown in Fig. 5.2.6. The heart of the arbiter is an 8b Manchester carry chain composed of two 4b chains. The equations for an 8b arbiter are: P[I] = ~(Req[I]+Token[I]), G[I] = Token[I], Grant[I] = C[I] & Req[I]. There is a single token present that represents the last processor to have received a grant. Because the propagate signal is enabled for processors not requesting the bus, the carry (token) will propagate left across the carry chain until encountering a request. Note that unlike a carry chain in an adder, the carry out of the arbiter wraps back to the carry in, potentially allowing a carry (token) to traverse the entire 8b. The 8b arbiter described above is easily extended to 16b using the same carry chain with minimal impact on the critical path. The arbitration requires 400ps with the remainder of the cycle time (1200ps) going to routing the requests and grants to/from the processors. The significant routing delay is minimized by locating the arbiter halfway between the two furthest processors.

The design methodology is a combination of industry-standard synthesis and place-and-route flow for control blocks, and full custom schematic/layout design for the datapath-style units. The core clock is generated by an on-chip all-digital PLL. Global clock distribution is full custom and consists of a power-efficient variable-density grid that minimizes total metal capacitance while maintaining low resistance paths to the heaviest clock loads. Global clock distribution power is <1W at 1.2V, 600MHz for a skew of <50ps. Local conditional clocks are two gain stages from the global clock and are designed on an ad-hoc basis. The entire chip is on the same clock domain except for interface clocks that are independently generated and distributed. Global floorplanning and wiring is done with an in-house tool that handles routing in addition to optimal repeater, local clock driver and decoupling capacitance placement. Timing closure for custom circuits is achieved with an in-house transistor-level analyzer. Industry-standard timing analyzers are used for global timing analysis.

First silicon is fully functional and successfully booted the OS.

Acknowledgements:
The authors acknowledge contributions of everyone in architecture, implementation, layout, verification, board design and software.

References:
[1] S. Kaneko, et al, "A 600MHz Single-Chip Multiprocessor with 4.8GB/s Internal Shared Pipelined Bus and 512kB Internal Memory," *ISSCC Dig. Tech. Papers*, pp. 254-255, Feb., 2003.

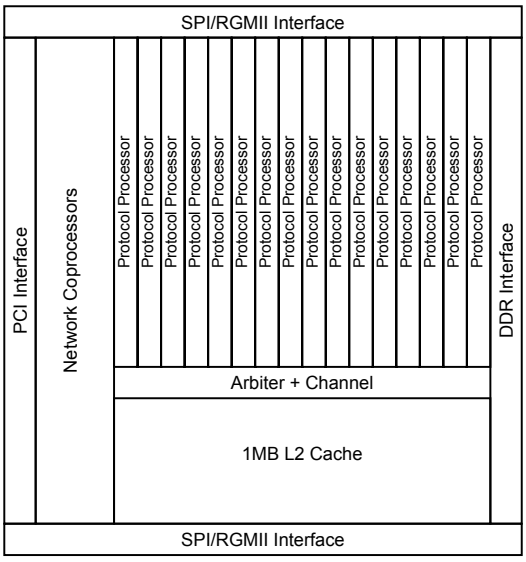


Figure 5.2.1: Chip floorplan.

Instructions per second	19.2 GOPS
Symmetric Cryptography	37Gb/s
RSA Operations (1024b private key)	16kops/s
Packet Processing	30M packets/s
String matching (DFA)	4Gb/s
TCP Connections	2M/s
TCP Termination Throughput	20Gb/s
ZIP Compression	4Gb/s
ZIP Decompression	4Gb/s
Main Memory Bandwidth	14.4GB/s

Figure 5.2.2: Peak performance.

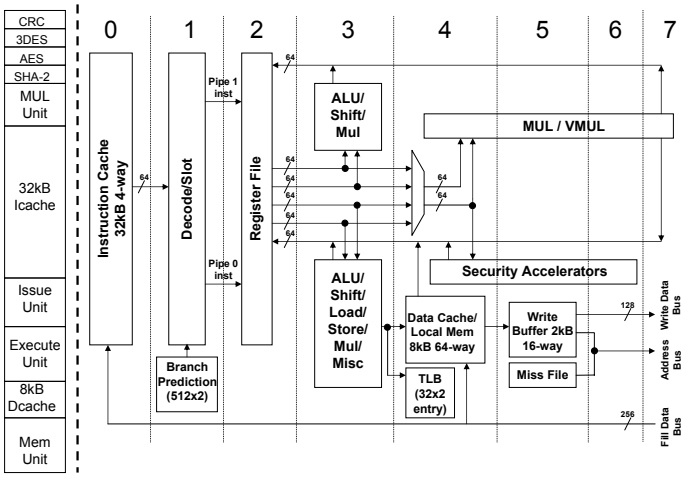


Figure 5.2.3: RISC processor.

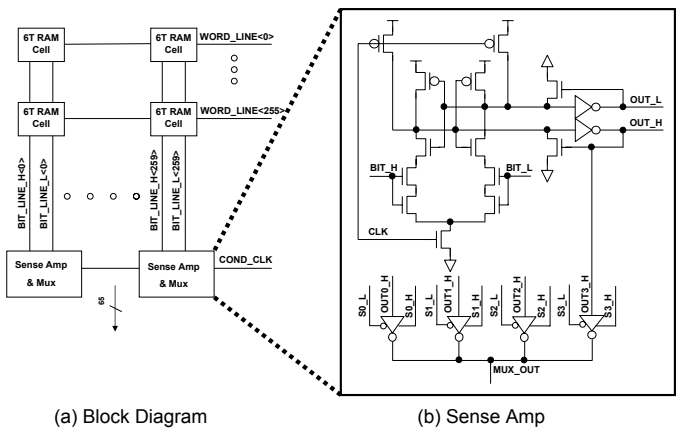


Figure 5.2.4: Icache.

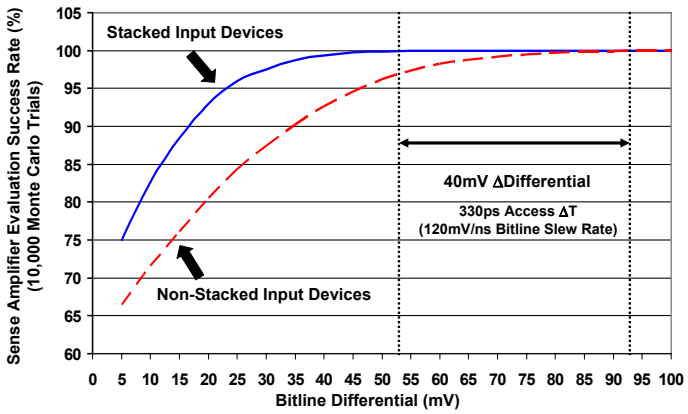
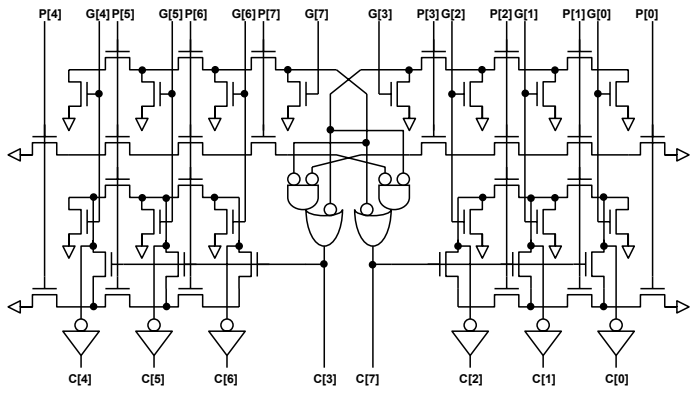


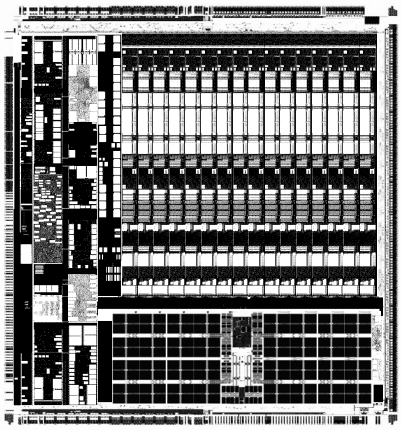
Figure 5.2.5: Sense-amplifier Monte-Carlo analysis.



Note: Precharge and keeper devices are not shown for clarity.

Figure 5.2.6: Arbiter.

Continued on Page 641



Number of transistors	180 million
Power	25W
Frequency	600MHz
Voltage	1.2V
Process	0.13μm CMOS
Number of Metal Layers	9
Number of MIPS Cores	16
Size of Icache (each core)	32kB
Size of Dcache (each core)	8kB
L2 Cache	1MB

Figure 5.2.7: Chip plot.

